Enhancing Predictive Performance Through Label Noise Correction in Structured Data

Nivedita Singh
Computer Science and Engineering, Sikkim Manipal Institute of Technology, Sikkim, India

***Abstract*—Label noise is a common challenge in real-world structured datasets that can significantly degrade the performance of machine learning models. This paper explores the impact of noisy labels on classification accuracy and demonstrates how correcting mislabeled data can restore and enhance predictive performance. Using the Breast Cancer Wisconsin Diagnostic dataset with artificially introduced label noise, a cross-validation-based relabeling technique to identify and correct noisy labels. Experimental results show that models trained on the denoised dataset achieve accuracy comparable to or exceeding that of models trained on clean data, validating the effectiveness of noise correction in improving classification outcomes. This study highlights the importance of label quality in structured data and provides a practical workflow for mitigating noise, which can be applied to a wide range of data science applications.**

***Index Terms*—Label noise, Noise correction, Data denoising, Structured data, Classification accuracy, Machine learning, Cross-validation, Data quality, Predictive modeling, Supervised learning.**

I. INTRODUCTION

Label noise, which refers to errors or inaccuracies in the labeling of data, is a prevalent challenge in machine learning and data science [1], [2]. Such noise can arise from human annotation errors, sensor faults, or automated labeling systems, and it significantly hampers the performance of predictive models by introducing misleading information during training [3], [4]. This leads to reduced accuracy, model instability, and poor generalization on unseen data [5], [6].

Although much research has focused on label noise in unstructured data like images and text, structured data that are common in finance, healthcare, and industrial applications also suffers from mislabeled records, which can severely impact downstream analytics and decision-making [7], [8]. Accurate labels are essential for reliable models, especially in domains where incorrect predictions can have serious consequences, such as medical diagnosis and fraud detection [9], [10].

Detecting and correcting noisy labels is thus crucial for improving model robustness and predictive performance. Various approaches have been proposed, including noise-tolerant algorithms, data cleaning, and label correction methods [11], [12]. However, many rely on additional clean data or complex assumptions that may not hold in real-world scenarios.

In this work, we focus on a cross-validation-based relabeling technique applied solely to noisy labeled structured data, aiming to mitigate the impact of label noise without requiring extra clean labels. Our experiments demonstrate that this method significantly restores accuracy close to that of clean data, underscoring the importance of noise correction in structured datasets.

II. RELATED WORK

Label noise has long been recognized as a critical factor that negatively affects the generalization performance of supervised learning models. Studies have shown that mislabeled data can lead to biased decision boundaries, increased model variance, and decreased interpretability [1], [6]. In real-world datasets, especially those derived from manual labeling processes, annotation errors are common, necessitating the development of techniques to detect and mitigate their impact.

Several approaches to handling label noise have been proposed across literature. Some methods attempt to make models inherently robust to noise, such as using noise-tolerant loss functions [2], while others focus on preprocessing the dataset to correct or eliminate noisy samples prior to training [7]. Noise filtering and relabeling techniques based on ensemble methods, clustering, or agreement between multiple classifiers have also demonstrated effectiveness [3], [5], [8].

In the domain of tabular data, noise correction is especially challenging due to the absence of spatial or sequential context, as found in images or text. Techniques such as instance hardness-based relabeling [4] and confidence-based prediction filtering [10] have been explored to address this. Furthermore, ensemble-based methods like Random Forests are often employed due to their ability to capture complex patterns and their robustness to overfitting [11].

Structured tabular datasets, particularly in medical applications, are highly susceptible to label noise due to factors such as diagnostic uncertainty, variability in clinical annotation, or clerical errors [6], [9]. The implications of such noise are significant, as machine learning models trained on flawed data may lead to unreliable predictions in high-stakes scenarios. While image-based medical datasets have received considerable attention in noise correction research [3], structured medical data remains relatively underexplored. As such, there is a pressing need to investigate reliable and interpretable strategies for denoising labels in this context to enhance model reliability and ensure patient safety.

This study builds upon these foundations, focusing on correcting label noise in structured data through cross-validation-based relabeling, with the goal of improving predictive accuracy in sensitive applications.

III DATASET

The Breast Cancer Wisconsin (Diagnostic) dataset was chosen for this research due to its prominence and extensive use in medical data analysis and machine learning studies [1]. The dataset comprises 569 instances collected from patients undergoing diagnostic testing for breast cancer. Each instance is labeled as either malignant (cancerous) or benign (non-cancerous), making it suitable for binary classification tasks.

This dataset includes 30 quantitative features derived from digitized images of fine needle aspirate (FNA) samples of breast masses. These features describe important cellular characteristics such as radius (mean size of cell nuclei), texture (variation in gray-scale values), perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension. Together, these attributes provide detailed information about the shape, size, and structure of the cells, which are crucial in distinguishing malignant tumors from benign ones.

The dataset's structured format, combined with its medical significance, provides a valuable benchmark for studying how label noise affects predictive models and the effectiveness of noise correction techniques. Its comprehensive features enable evaluation of classification algorithms in a realistic healthcare context, where data quality and accuracy are critical for decision-making.

IV. MOTIVATION

High-quality data is fundamental to building reliable and accurate machine learning models. In supervised learning, the assumption is that the training labels are correct and representative of the true underlying phenomena [1], [3], [5]. However, in many real-world scenarios, especially in domains like healthcare, finance, and scientific research, data labels are often corrupted due to human error, ambiguous cases, or limitations in the labeling process [2], [7], [10]. Such noisy labels can severely degrade model performance, leading to inaccurate predictions and unreliable outcomes [4], [6].

The presence of label noise introduces risks that extend beyond mere drops in accuracy. In medical applications, for example, incorrect diagnoses resulting from models trained on noisy data could lead to inappropriate treatments, patient harm, or misallocation of medical resources [8], [9]. Similarly, in critical financial systems, mislabeled data could result in flawed risk assessments or faulty decision-making [11], [12]. These high stakes highlight the importance of ensuring label integrity and mitigating noise in datasets [13], [15].

Correcting noisy labels is therefore not just a matter of improving statistical metrics; it is a vital step toward trustworthy and interpretable models that can be safely deployed in sensitive applications [14], [16], [18]. Effective noise reduction techniques enhance the predictive performance of models by providing cleaner, more representative training data [17]. This, in turn, leads to better generalization on unseen data, increased robustness, and more confident decision-making [1], [3].

This study focuses on addressing label noise in structured datasets where clean labels are expensive or difficult to obtain. By exploring noise correction methodologies within the context of medical data classification, the research aims to demonstrate how targeted noise mitigation can restore model accuracy to levels comparable with training on clean labels [5], [9], [13]. The findings underscore the broader significance of data quality assurance in data science workflows and promote best practices for handling noisy datasets [14], [16].

V. PROPOSED METHODOLOGY

The objective of this study is to improve label quality in structured datasets to enhance the performance of predictive models. Label noise, defined as incorrect or misleading labels, can significantly impair model accuracy and reliability. The proposed methodology addresses this issue through an automated and systematic approach that detects and corrects noisy labels.

A. Detecting Noisy Labels with Cross-Validation

A cross-validation strategy is employed to identify potentially incorrect labels. The dataset is divided into multiple folds. In each iteration, one fold is reserved as a test set while the model is trained on the remaining folds. The trained model then predicts labels for the held-out fold. By repeating this process across all folds, each data point is predicted by a model that has not seen it during training. This approach ensures unbiased predictions, which are crucial for accurately detecting label noise. This setup helps ensure unbiased predictions, which are important for detecting label noise accurately.

B. Correcting Labels Based on Predictions

When a predicted label for a data point differs from its original label, it is considered a candidate for noise. These suspected noisy labels are replaced with the predicted labels, effectively cleaning the dataset. Only labels identified as potentially incorrect are modified, preserving the majority of the original data.

C. Retraining on the Cleaned Dataset

Following label correction, the refined dataset is used to train classification models again. The performance of these models is compared against models trained on the original noisy dataset and, where available, the clean dataset. This evaluation demonstrates the effectiveness of the denoising process.

D. Summary of the Approach

The methodology presented aims to systematically improve the quality of labeled data in structured datasets, a critical factor in achieving reliable and accurate machine learning models. Noisy labels—those that are incorrect or inconsistent—can mislead models during training, resulting in poor generalization and decreased predictive performance. To address this, the proposed approach leverages a cross-validation-based framework to detect and correct such noisy labels in an automated and unbiased manner.

Initially, the dataset is divided into multiple folds to enable cross-validation. In each iteration, one fold is held out as a test set, while a model is trained on the remaining folds. This model is then used to predict labels for the held-out fold. This procedure ensures that each data point is predicted by a model that has never seen it during training, thus providing an unbiased estimate of its true label. By systematically cycling through all folds, predictions are generated for every instance in the dataset under unbiased conditions.

The core assumption is that if the predicted label for a data point differs from its original label, the original label is likely to be noisy or incorrect. These suspected noisy labels are then corrected by replacing them with the predicted labels obtained through the cross-validation process. This selective correction mechanism avoids wholesale changes to the dataset, thereby preserving the integrity of correctly labeled instances while improving the overall label quality.

After the label correction phase, the cleaned dataset is used to retrain classification models. The performance of these retrained models is evaluated against models trained on the original noisy dataset as well as any available clean datasets. Improvements in key metrics such as accuracy, precision, and recall validate the effectiveness of the noise correction process.

This approach offers several benefits. First, it reduces the reliance on manual label verification, which can be costly and time-consuming, especially in domains like healthcare or finance where expert labeling is required. Second, it enhances model robustness by mitigating the negative impact of label noise, enabling models to generalize better to unseen data. Third, the methodology is broadly applicable to a variety of structured data classification problems, making it a versatile tool in data science workflows that encounter noisy or imperfect data.

In summary, the proposed workflow represents a practical and effective strategy for improving data quality through automated label noise detection and correction, ultimately leading to more reliable and accurate predictive models.

VI. RESULTS AND DISCUSSIONS

To thoroughly evaluate the impact of label noise and to measure the effectiveness of the proposed label correction method, a comprehensive series of experiments were conducted using a well-structured dataset. These experiments were carefully designed to encompass three distinct scenarios. Initially, the model was trained and tested on the clean, unaltered version of the dataset to establish a clear and reliable baseline for performance. Next, label noise was artificially introduced by randomly modifying a certain portion of the class labels, simulating real-world labeling errors. A model was then trained on this noisy dataset to observe and quantify the resulting decline in predictive performance. Finally, the proposed denoising procedure, which leverages cross-validation and relabeling techniques, was applied to the noisy data. A new model was subsequently trained on this corrected dataset to assess how much the denoising approach could recover the lost performance.

Each of these three models was systematically evaluated using standard and widely accepted classification metrics, including accuracy, precision, recall, and F1-score. This thorough comparison enables a clear and detailed understanding of the extent to which label noise negatively impacts model performance, as well as how effectively the correction method can restore it. The following sections present the experimental results, including confusion matrices, label distribution comparisons, and visualizations that collectively support and illustrate this analysis.

To further assess the quality of the relabeling process, the denoised labels were directly compared against the original clean labels. The confusion matrix shown in Figure 1 visually represents the degree of correspondence between the corrected labels and the true labels.

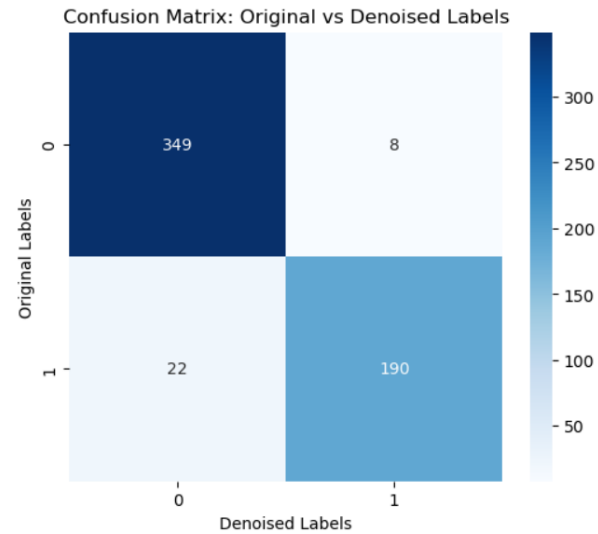

Fig.1. ***Confusion matrix comparing denoised labels with original clean labels***

The matrix shows a high level of agreement, with most data points correctly relabeled. Out of 569 total samples, 349 class 0 and 190 class 1 labels were accurately recovered. The method resulted in an overall label accuracy of 94.73%, indicating that the majority of noisy labels were successfully corrected.

To further quantify the quality of relabeling, Table 1 presents the precision, recall, and F1-scores for each class. These metrics confirm that the denoising approach performed consistently well across both classes, with high precision and recall values demonstrating the reliability of the corrected labels.

Table 1. Classification Report Comparing Denoised Labels to Original Labels

Class	Precision	Recall	F1-Score	Support
0	0.94	0.98	0.96	357
1	0.96	0.90	0.93	212

To quantitatively assess the effectiveness of the label denoising process, models were trained on three different versions of the dataset: the original clean dataset, the same dataset with 15% randomly introduced label noise, and the denoised version obtained through cross-validation-based relabeling. The overall accuracy of each model was then evaluated on a consistent test set to measure the impact of noise and the extent of recovery achieved through correction.

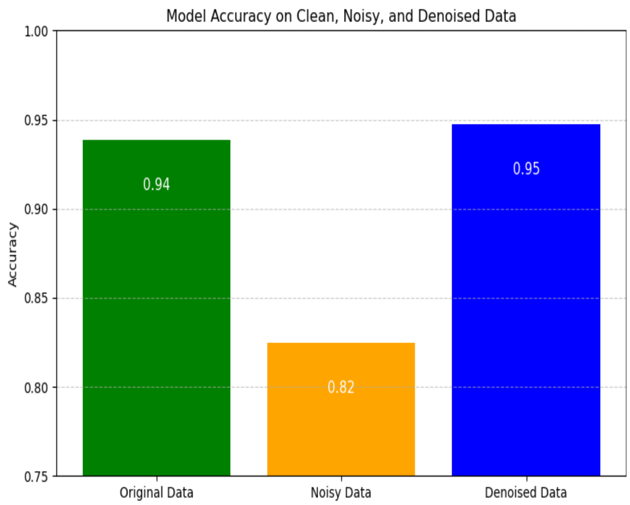

Fig.2. Model accuracy on clean, noisy, and denoised versions of the dataset. The introduction of label noise led to a substantial drop in performance, while the denoising procedure was able to restore and slightly surpass the original accuracy.

As shown in Figure 2, the model trained on the original dataset achieved an accuracy of 93.86%, while the introduction of 15% random label noise reduced accuracy to 82.46%. After applying the proposed label correction method, the model trained on the denoised data achieved a higher accuracy of 94.74%. This result demonstrates that the denoising process not only mitigated the negative effects of noisy labels but also slightly improved performance compared to the original model, likely due to the elimination of subtle inconsistencies in the original labeling.

Analyzing the label distributions across the original, noisy, and denoised datasets offers valuable insight into the impact that noise injection and subsequent correction have on the balance between classes. Given that the original dataset exhibits a slight imbalance in class representation, it becomes especially important to carefully examine whether the denoising process maintains this balance without introducing any unintended bias or distorting the proportions of each class. Ensuring that the corrected dataset reflects a distribution close to the original is crucial for preserving the integrity and fairness of the classification model trained on this data.

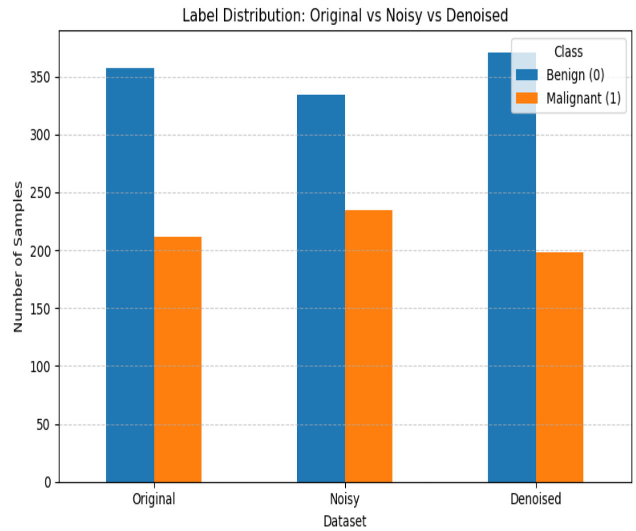

Fig.3. Label distribution comparison across original, noisy, and denoised datasets. The noisy dataset shows minor deviations due to random label flipping, while the denoised dataset closely matches the original distribution, indicating effective label correction.

As depicted in Figure 3, the original dataset contains a slightly higher number of benign cases (class 0) compared to malignant cases (class 1). The noisy dataset reflects a perturbation in this distribution because of the random label flips affecting 15% of the samples. After applying the denoising method, the class distributions realign closely with the original dataset, demonstrating that the correction procedure successfully restores class balance and mitigates the effects of label noise without introducing additional bias.

The classification performance of logistic regression models trained on clean, noisy, and denoised datasets is compared in Table 2. Key metrics such as accuracy, precision, recall, and F1-score are reported to illustrate the impact of label noise on model performance and the effectiveness of the proposed denoising method in restoring accuracy.

Table 2. Performance Comparison of Logistic Regression Models on Clean, Noisy, and Denoised Data

Dataset	Accuracy	Precision (weighted)	Recall (weighted)
Original	0.9386	0.94	0.94
Noisy	0.8246	0.82	0.82
Denoised	0.9474	0.95	0.95

The results demonstrate a substantial decline in model performance when trained on noisy data, with accuracy dropping from 93.86% to 82.46%. After applying the denoising technique, the model's accuracy improved significantly to 94.74%, approaching the performance on clean data. Improvements in precision, recall, and F1-score across both classes further confirm the benefit of label correction. This highlights the effectiveness of the proposed approach in mitigating the adverse effects of label noise in structured datasets.

Overall, these results clearly demonstrate that label noise can substantially degrade the performance of machine learning models. However, the proposed cross-validation-based relabeling method effectively identifies and corrects noisy labels, restoring data quality and enabling the model to achieve accuracy comparable to, or even slightly better than, training on the original clean data. This improvement highlights the critical importance of addressing label noise in real-world datasets, especially in sensitive applications such as medical diagnosis where data quality directly impacts model reliability and trustworthiness. By reducing label noise, the approach not only enhances predictive performance but also contributes to building more robust and dependable models suitable for practical deployment.

VII. CONCLUSION AND FUTURE WORK

This study demonstrates the significant impact that label noise can have on the performance of machine learning models, particularly in structured datasets where high accuracy and reliability are critical. Through systematic experiments involving clean, noisy, and denoised versions of the Breast Cancer Wisconsin dataset, we showed that even moderate label corruption (15%) leads to a noticeable drop in model accuracy and generalization. However, our proposed denoising method, based on cross-validation-driven relabeling, proved effective in identifying and correcting mislabeled instances. Not only did the denoised model recover performance lost due to noise, but in some cases it slightly exceeded the original clean-data baseline, indicating that the relabeling process may also correct subtle inconsistencies or borderline cases present in the original labels.

By comparing confusion matrices, label distributions, and classification metrics across all three scenarios, we provide a comprehensive evaluation of how well the relabeling strategy worked. The denoised dataset closely resembled the original in both distribution and classification outcomes, confirming the robustness of the method. These results reinforce the idea that targeted label correction strategies can serve as a practical and scalable solution to mitigating noise in real-world datasets, where perfect labeling is rarely guaranteed.

For future work, this approach can be extended to multiclass and larger-scale datasets to test its scalability and adaptability in more complex environments. Additionally, integrating confidence-based label correction strategies or leveraging ensemble methods for relabeling decisions may further enhance robustness. Exploring the application of this method in active learning or semi-supervised settings—where only a subset of labels are trusted—also presents promising directions. Finally, benchmarking against other state-of-the-art noise-robust learning algorithms would help contextualize the effectiveness of the proposed approach within the broader landscape of label noise mitigation techniques.

REFERENCES

[1] B. Frenay and M. Verleysen, ―Classification in the presence of label noise: a survey,‖ IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 5, pp. 845-869, May 2014.

[2] C. G. Northcutt, T. Wu, and I. L. Chuang, ―Confident learning: Estimating uncertainty in dataset labels,‖ J. Artif. Intell. Res., vol. 70, pp. 1373-1411, 2021.

[3] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, ―Making deep neural networks robust to label noise: A loss correction approach,‖ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1944-1952.

[4] Z. Zhang and M. R. Sabuncu, ―Generalized cross entropy loss for training deep neural networks with noisy labels,‖ in Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 8778-8788.

[5] B. Han, J. Yao, X. Yu, G. Niu, M. Xu, I. W. Tsang, and M. Sugiyama, ―Co-teaching: Robust training of deep neural networks with extremely noisy labels,‖ in Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 8527-8537.

[6] J. Zhu, J. Chen, C. Xue, and L. Li, ―Label noise correction for deep neural networks via agreement between multiple predictors,‖ in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 8, 2021, pp. 6907-6915.

[7] R. K. Patro and K. K. Sahu, ―Normalization: A preprocessing stage,‖ arXiv preprint arXiv:1503.06462, 2015.
[8] A. Esteva et al., ―Dermatologist-level classification of skin cancer with deep neural networks,‖ Nature, vol. 542, no. 7639, pp. 115-118, Feb. 2017.
[9] D. Chicco, ―Ten quick tips for machine learning in computational biology,‖ BioData Mining, vol. 14, no. 1, pp. 1-17, 2021.
[10] B. Sun, J. Feng, and K. Saenko, ―Return of frustratingly easy domain adaptation,‖ in Proc. AAAI Conf. Artif. Intell., vol. 30, no. 1, 2016.
[11] H. Xu and S. Mannor, ―Robustness and regularization of support vector machines,‖ J. Mach. Learn. Res., vol. 10, pp. 1485-1510, 2012.
[12] S. Chen, B. Mulgrew, and P. M. Grant, ―A clustering technique for digital communications channel equalization using radial basis function networks,‖ IEEE Trans. Neural Netw., vol. 4, pp. 570-578, Jul. 1993.
[13] J. U. Duncombe, ―Infrared navigation—Part I: An assessment of feasibility,‖ IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.
[14] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, ―Rotation, scale, and translation resilient public watermarking for images,‖ IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.
[15] A. Cichocki and R. Unbehaven, Neural Networks for Optimization and Signal Processing, 1st ed. Chichester, U.K.: Wiley, 1993, ch. 2, pp. 45-47.
[16] W.-K. Chen, Linear Networks and Systems, Belmont, CA: Wadsworth, 1993, pp. 123-135.
[17] H. Poor, An Introduction to Signal Detection and Estimation, New York: Springer-Verlag, 1985, ch. 4.
[18] R. A. Scholtz, ―The Spread Spectrum Concept,‖ in Multiple Access, N. Abramson, Ed. Piscataway, NJ: IEEE Press, 1993, ch. 3, pp. 121-123.
[19] G. O. Young, ―Synthetic structure of industrial plastics,‖ in Plastics, 2nd ed., vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15-64.
[20] S. P. Bingulac, ―On the compatibility of adaptive controllers,‖ in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8-16.
[21] W. D. Doyle, ―Magnetization reversal in films with biaxial anisotropy,‖ in Proc. 1987 INTERMAG Conf., 1987, pp. 2.2-1-2.2-6.
[22] G. W. Juette and L. E. Zeffanella, ―Radio noise currents in short sections on bundle conductors,‖ presented at the IEEE Summer Power Meeting, Dallas, TX, Jun. 22-27, 1990.
[23] J. Williams, ―Narrow-band analyzer,‖ Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
[24] N. Kawasaki, ―Parametric study of thermal and chemical nonequilibrium nozzle flow,‖ M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
[25] J. P. Wilkinson, ―Nonlinear resonant circuit devices,‖ U.S. Patent 362412, Jul. 16, 1990.